Sept 20th

Feature space - set of all possible data points.

COMPARISON REQUIRES DISSIMILARITY FUNC [input 2 data points, large value → v. dissimilar]

- one instance is a distance func.

MINKOWSKI

ALSO CAN USE SIMILARITY FUNC [large value → v. similar]

COSINE SIMILARITY

$S(x,y) = \cos\theta$ where θ is angle b/w x, y

WE CAN GET CORRESPONDING DISSIM FUN

$d(x,y) = \frac{1}{S(x,y)}$ OR $d(x,y) = K - S(x,y)$ for some K.

$d(x,y) = 1 - S(x,y)$

WE USE COSINE OVER EUCLIDEAN DISTANCE? → WHEN DIRECTION MATTERS MORE THAN MAGNITUDE

Jaccard Similarity

ex: take text and each word is an attribute, and values are whether it is in the given doc.

→ could use MANHATTAN → Returns size of set difference.
PROBLEM IS IT CAN'T DISTINGUISH HOW MANY HAD THE SAME VALUE.
WE WANT SOME NOTION OF INTERSECTION

$JSim(x,y) = \frac{|x \cap y|}{|x \cup y|}$ $JDist(x,y) = 1 - \frac{|x \cap y|}{|x \cup y|}$